CLAIMS

What is claimed is:

1. A method for clustering data in a system having an integrator and at least two computing units comprising:

(a) loading each computing unit with common global parameter values and a particular local data set;

(b) each computing unit generating local sufficient statistics based on the local data set and global parameter values; and

(c) employing the local sufficient statistics of all the computing units to update the global parameter values.

2. The method of claim 1 wherein the step of loading each computing unit with common global parameter values and a particular local data set further comprises:

a_1) receiving a set of data points to be clustered;

a_2) dividing the data points into at least two local data sets;

a_3) sending common global parameter values to each of the computing units; and

a_4) sending each local data sets to a designated computing unit.

3. The method of claim 2 wherein the step of employing the local sufficient statistics of all the computing units to update the global parameter values further comprises:

c_1) each computing unit sending its local sufficient statistics to the integrator;

c_2) the integrator determining global sufficient statistics based on the local sufficient statistics of all the computing units; and

c_3) the integrator determining updated global parameter values based on the global sufficient statistics.

4. The method of claim 1 further comprising:

d) checking the convergence quality;

e) determining whether the convergence meets a predetermined quality; and

f) when the convergence meets a predetermined quality, stop processing; otherwise;

g) when the convergence fails to meet a predetermined quality, providing the updated global parameter values to the computing units and repeating steps (a) to (c).

5. The method of claim 2 wherein sending common global parameter values to each of the computing units includes the step of:

broadcasting common global parameter values to each of the computing units.

6. The method of claim 2 further comprising the step of:

initializing the common global parameter values before sending the common global parameter values to each of the computing units.

7. The method of claim 1 wherein a distributed K-Means clustering algorithm is implemented.

8. The method of claim 1 wherein a distributed K-Harmonic Means clustering algorithm is implemented.

9. The method of claim 1 wherein a distributed Expectation-Maximization (EM) clustering algorithm is implemented.

10. The method of claim 1 wherein the data points to be clustered are naturally distributed.

11. A distributed data clustering system comprising:

(a) a first computing unit for performing data clustering based on a first local data set that is a subset of data points to be clustered and global parameter values to generate first local sufficient statistics;

(b) a second computing unit for performing data clustering based on a second local data set that is a subset of the data points to be clustered and global parameter values to generate second local sufficient statistics; and

(c) a integrator unit for receiving the first and second local sufficient statistics from the first and second computing units, respectively, and for employing the first and second local sufficient statistics to update the global parameter values.

12. The distributed data clustering system of claim 11 wherein the system performs a distributed K-Means clustering algorithm.

13. The distributed data clustering system of claim 11 wherein the system performs a distributed K-Harmonic Means clustering algorithm.

14. The distributed data clustering system of claim 11 wherein the system performs a distributed Expectation-Maximization (EM) clustering algorithm.

15. The distributed data clustering system of claim 11 wherein the first and second local data sets include data points that are naturally distributed.

16. The distributed data clustering system of claim 11 wherein the integrator receives a set of data points to be clustered, divides the data points into at least two local data sets, sends common global parameter values to each of the computing units, and sends each of the local data sets to a designated computing unit.

17. The distributed data clustering system of claim 11 wherein the integrator receives the local sufficient statistics from the first and second computing units; and

wherein the integrator determining global sufficient statistics based on the local sufficient statistics of the first and second computing units; and

wherein the integrator determines updated global parameter values based on the global sufficient statistics.

18. The distributed data clustering system of claim 11 wherein the integrator checks the convergence quality, determines whether the convergence meets a

predetermined quality, and when the convergence meets a predetermined quality, the integrator stops processing, when the convergence fails to meet a predetermined quality, the integrator provides updated global parameter values to the computing units.

19. The distributed data clustering system of claim 11 wherein the integrator broadcasts common global parameter values to the first and second computing units.

20. The distributed data clustering system of claim 11 wherein the integrator initializes the common global parameter values before sending the common global parameter values to the first and second computing units.